Improving settlement selection for small-scale maps using data enrichment and machine learning

Karsznia, Izabela ; Weibel, Robert

Abstract: Acquiring and formalizing cartographic knowledge still is a challenge, especially when the generalization process concerns small-scale maps. We concentrate on the settlement selection process for small-scale maps, with the aim of rendering it more holistic, and making methodological contributions in four areas. First, we show how written specifications and rules can be validated against the actual published map products, thus pointing to gaps and potential improvements. Second, we use data enrichment based on supplementing information extracted from point-of-interest data in order to assign functional importance to particular settlements. Third, we use machine learning (ML) algorithms to infer additional rules from existing maps, thus making explicit the deep knowledge of cartographers and allowing to extend the cartographic rule set. And fourth, we show how the results of ML can be transformed into human-readable form for potential use in the guidelines of national mapping agencies. We use the case of settlement selection in the small-scale maps published by the Polish national mapping agency (GUGiK). However, we believe that the methods and findings of this paper can be adapted to other environments with minor modifications.

DOI: https://doi.org/10.1080/15230406.2016.1274237

Improving Settlement Selection for Small-scale Maps Using Data Enrichment and Machine Learning

Izabela Karsznia[*], Robert Weibel[**]

[*]*Department of Geoinformatics, Cartography and Remote Sensing, University of Warsaw, Krakowskie Przedmieście 30, 00-927 Warsaw, Poland*

Corresponding author: i.karsznia@uw.edu.pl

[**]*Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

Improving Settlement Selection for Small-scale Maps Using Data Enrichment and Machine Learning

Acquiring and formalizing cartographic knowledge still is a challenge, especially when the generalization process concerns small-scale maps. We concentrate on the settlement selection process for small-scale maps, with the aim of rendering it more holistic, and making methodological contributions in four areas. First, we show how written specifications and rules can be validated against the actual published map products, thus pointing to gaps and potential improvements. Second, we use data enrichment based on supplementing information extracted from point-of-interest data in order to assign functional importance to particular settlements. Third, we use machine learning algorithms to infer additional rules from existing maps, thus making explicit the deep knowledge of cartographers and allowing to extend the cartographic rule set. And fourth, we show how the results of machine learning can be transformed into human-readable form for potential use in the guidelines of national mapping agencies. We use the case of settlement selection in the small-scale maps published by the Polish national mapping agency (GUGiK). However, we believe that the methods and findings of this paper can be adapted to other environments with minor modifications.

Keywords: settlement selection; small-scale mapping; data enrichment; machine learning.

Introduction

Research in map generalization since its beginnings has mainly concentrated on topographic maps and databases at large to medium scales (e.g. Brassel and Weibel 1988; Müller et al. 1995; Mackaness et al., 2007; Burghardt et al., 2014; Stoter et al. 2009, 2010, 2014). On the other hand, the advent of multi-resolution databases (MRDBs) makes a direct case for the development of medium and small-scale map generalization procedures and algorithms. Having such generalization techniques available is of interest not only to national mapping agencies (NMAs) in the production of topographic maps and MRDBs, but also to scientists exploring their data at regional

scales, or commercial companies developing systems supporting the generalization process. However, although some solutions for small-scale mapping have been proposed (West-Nielsen and Meyer 2007; Samsonov and Krivosheina 2012), the generalization toolbox for small scales is still only sparsely filled, and we are still a long way from a comprehensive and formalized methodology for small-scale generalization.

This paper addresses the problem of automated settlement selection in small-scale maps in the range of 1:250 000 to 1:500 000, with a focus on two processes of knowledge formalization: first, the extraction of semantic and structural knowledge (data enrichment), and second, the automated acquisition and utilization of procedural cartographic knowledge by machine learning (ML). This work thus contributes to extending the toolbox for small-scale mapping.

We use the case of settlement selection for the small-scale maps published by the Polish national mapping agency (GUGiK). However, since the situation regarding topographic databases and map products is fairly similar across different national mapping agencies (NMAs), despite cultural differences and preferences (Stoter et al., 2010; Duchêne et al., 2014), we trust that the contributions and findings of this paper can be easily adapted to the context of other small-scale map production processes. In summary, the paper makes methodological contributions in four different areas:

First, like other NMAs, GUGiK has guidelines for small-scale map generalization. Similarly to the situation at other NMAs these written specifications and rules have not been experimentally validated so far, however. We show how written specifications and rules can be validated against the actual published map products, assessing the completeness and accuracy of the rule set, and pointing to gaps and potential improvements.

Second, in our work we have been using GUGiK's General Geographic Database (GGD) at the nominal scale of 1:250 000 as a source database for small-scale map generalization. The quantitative analysis of settlements revealed that the content of this source database was semantically rather poor, with few attributes, a fact that made it hard to formalize rich and expressive generalization rules. Again, this semantic scarcity is also common in topographic database products of other NMAs, a fact that has given rise to a stream of research on data enrichment in cartography (Brassel and Weibel, 1998; Weibel; 1995; Neun et al., 2008, Stoter et al., 2014; Mackaness et al., 2014ab). We show how additional building-related attributes can be extracted from point-of-interest (POI) data and merged to semantically enrich the topographic source database, thus allowing to infer additional rules of settlement selection that take into account settlement importance.

Third, in manual map production, cartographers used rich knowledge entrusted to them in intensive training, as well as long-standing professional experience when crafting maps. This type of knowledge has been termed 'deep knowledge' by Muller and Mouwes (1990). It has been observed by many authors (Buttenfield and McMaster, 1991; Weibel, 1995; Kilpeläinen, 2000; Duchêne et al., 2005; Harrie and Weibel, 2007; Balley et al., 2014) that it is by no means a trivial task to make explicit and formalize this deep cartographic knowledge. In this paper, we show how ML algorithms can be used to learn additional rules from existing maps, thus making explicit the deep knowledge of cartographers and significantly improving the completeness and accuracy of the cartographic rule set.

Fourth, the usage of ML techniques proved to be very successful in revealing additional rules for settlement selection in small-scale maps. In this respect, we show how by using appropriate ML techniques — decision trees in our case — rules can be

generated in a human-readable form from the output of ML, such that they can be used to extend the written guidelines of NMAs and thus directly contribute to cartographic practice.

Related work

Settlement selection algorithms

Selection of relevant map objects is the first and in many ways the most important task that is undertaken during the map generalization process (Karsznia, 2013). This generalization operation mainly affects the quantity of visual information as it entails removing part of the map content (Stanislawski et al., 2014).

The well-known 'Radical Law' (Töpfer and Pillewizer, 1966) was and still is extensively adopted by other authors for the purpose of object selection or quantitative evaluation of generalization results. However, it does not follow from this equation which objects exactly should be retained; the problem of qualitative selection is not addressed (Stanislawski et al., 2014). This could be remedied by supplying a ranked list of map objects to the algorithm (e.g. settlements ranked by population). But even then, the problem persists that the Radical Law does not take into account local variations in the density of map objects (Sarjakoski, 2007).

A thorough review of selection algorithms available in the cartographic literature has been conducted by Li (2007), dividing selection algorithms into two groups, algorithms for so-called selective omission of point objects, and algorithms for the simplification of the structure of a set of point objects.

Within the first group Langran and Poiker (1986) presented five early algorithms for settlement selection, most of which relied on some form of ranking, some also taking into account the density of map objects. Ranking was also used in the approach

proposed by Flewelling and Egenhofer (1993). Finally, Van Kreveld et al. (1997) presented an algorithm for settlement selection in interactive visualization that solves spatial conflicts arising due to object density, and maintains the spatial distribution characteristics of the generalized objects.

The second group of selection algorithms was designed to simplify the structure of the set of point objects based on a set of parameters describing it (Li, 2007). This operation has also been termed typification by other authors (Stanislawski et al., 2014). These algorithms must take some attribute and spatial characteristics of the point set into consideration, including: the weight of each point object (e.g. by population), spatial relationships among points (e.g. proximity, alignments), semantic relations (e.g. hierarchies), the spatial distribution of objects as well as their density. This group includes algorithms taking into account geometric object characteristics (Ai and Liu 2004) and algorithms considering both geometric and thematic information (Yan and Weibel, 2008; Samsonov and Krivosheina, 2012).

The main limitation of the existing algorithms is that they usually only consider very basic semantic attributes, such as the population and administrative status of the settlements. We argue that other semantic attributes should be considered to fully reflect settlement importance, augmented by measures describing the spatial context, such as local density differentiation and settlement size structure. As manually produced paper maps as well as the literature in urban geography show (Carol, 1960; Smith, 1965; Batty, 2006), other factors should also be considered, such as the settlements' touristic, cultural or education functions. While such attempts were reported in the early literature (Dixon, 1967; Kadmon, 1972; Richardson and Muller, 1991), the technology was not mature enough at that time to fully take advantage of such solutions.

Data enrichment

Spatial databases are typically rich in geometry, but rather poor in semantics. This is not surprising, as semantics is usually dependent on the context of usage; a building takes a different notion in an application for tax assessment than in an application for cultural heritage preservation. Spatial databases, however, in particular topographic databases, are general-purpose databases. On the other hand, generalization needs rich semantics (Mackaness, 2006). It is thus necessary to enrich the source database with additional important information concerning the special semantics of map objects and the relations between them. This process is commonly referred to as *data enrichment* (Neun, 2007; Mackaness et al., 2014ab; Stoter et al. 2014) and the process of analyzing groups of objects and relationships among them is often called the *structure recognition* process (Brassel and Weibel, 1988; Steiniger and Weibel 2007).

Data enrichment is also a common and useful practice applied in NMAs. For the purpose of deriving databases and maps at reduced detail NMAs use data enrichment processes to supply MRDBs with additional information used in generalization process (Bobzien et al., 2008). This supplementary information may have semantic (e.g. statistical) or geometric character (Stoter et al., 2014).

Formalization of Cartographic Knowledge

Cartographic knowledge is often not readily available as rules or other formal representations, thus necessitating significant efforts to acquire and formalize knowledge for usage in the automated generalization process. Three types of cartographic knowledge used during the generalization process have been identified (Armstrong 1991; Weibel 1995). *Geometrical knowledge* helps describing the geometry of generalized objects, such as its position, shape or distribution. *Structural knowledge*

contributes information about the overall object structure, and the geomorphological, economic or cultural meaning of map objects. Thus, it also relates to the semantics of objects (and thus is also termed semantic knowledge by some authors; Chang and Macmaster, 1993; Kulik et al., 2005; Dutton and Edwardes, 2006). Finally, *procedural knowledge* describes the knowledge about appropriate generalization operators that should be applied to the map objects under consideration, as well as the order in which these operations should be processed.

The geometric and structural knowledge is usually generated in the structure recognition process (Brassel and Weibel 1988) mentioned above, that is, a process of spatial analysis and pattern recognition leading to data enrichment. The procedural knowledge, however, rests with the expert cartographers and thus first has to be made explicit. Procedural knowledge about generalization can be acquired from four different sources (Weibel 1995; Stoter et al., 2014).

First, where they exist, textbooks and official NMA guidelines often serve as the primary source of generalization rules. This source was for instance used by Stoter et al. (2009; 2014) as a starting point for specifications to drive the generalization of large- to medium-scale maps. However, written specifications most often do not describe the generalization process entirely (Muller and Mouwes, 1990), necessitating further rule formation from other sources (Stoter et al. 2009; 2014).

A second source of cartographic knowledge is contributed by conventional knowledge acquisition through interviews with cartographic experts or direct observation of cartographers on the job. Early examples of this approach include Kilpeläinen (2000) for rule formation in topographic map generalization and Richardson and Muller (1991) in thematic atlas mapping. More recently, several projects have also started involving end users, besides cartographers, to evaluate map

products generated by an initial set of map specifications, thus enriching the set of generalization rules used (Stoter et al., 2009; Stoter et al., 2014; Balley et al. 2014; Duchêne et al. 2014).

The analysis of existing map series is a further source of cartographic knowledge. This approach is also called 'reverse engineering' (Weibel, 1995), as the knowledge acquisition process works its way back from the final map products at different scales. Leitner and Buttenfield (1995) used this approach for knowledge acquisition by analysis of the Austrian topographic map series, while Stoter et al. (2009) did the same for maps of the Netherlands.

Finally, machine learning has recently gained importance as a source of cartographic knowledge, starting off from early attempts in the 1990s (Weibel et al., 1995a; Weibel et al., 1995b; Plazanet et al., 1998). The main role and potential of ML techniques in knowledge acquisition for generalization is to find patterns in and extract generalization rules from large sets of empirical observations, a process that would be (too) hard to achieve for a human interpreter. The empirical observations required to drive ML may originate from the logging of interactions of expert users with an interactive cartography system (Weibel et al., 1995a; Duchêne et al., 2005). Alternatively, Ruas et al. (2006) have learned rules from logs of a self-evaluating generalization system to improve its efficiency. Empirical data may also originate from spatial analysis (Steiniger et al., 2010) and manually be tagged by expert users (Mustière et al., 2000). Machine learning may be used in support of generalization for two purposes: to generate an initial set of rules, when no previously formalized knowledge exists (Weibel et al., 1995a; Plazanet et al., 1998) or – which is the more frequent case – to extend an initial rule set by evaluating the performance of an existing

system (Mustière et al., 2000; Duchêne et al., 2005; Ruas et al., 2006; Sheeren et al., 2009).

Research methodology

Data

In this research two Polish national databases produced by GUGiK were used, the General Geographic Database (GGD) [1] and the Topographic Objects Database (BDOT10k). GGD has a nominal scale of 1:250 000 and serves as a starting point for the small-scale generalization process at GUGiK. In this work, the aim was to select settlements at 1:500 000. The database consists of eight thematic layers: administrative zoning, settlement and anthropogenic objects, hydrography, topography, transport, land cover and land use, protected and restricted areas, geographical names. This study focused on the *settlements* layer.

BDOT10k is the second database used and corresponds to a nominal scale of 1:10 000. It was used to enrich the information available in the GGD database (see Section **Data enrichment** below). The database includes twelve classes of topographic objects: buildings, infrastructure and equipment, address points, transportation network, watercourse network, land cover, land use, protected areas, public utilities, administrative division units, geodetic control networks, and other objects representing specific information about the topography. For the purpose of enriching the GGD settlements layer information from the *buildings* layer of BDOT10k was used to extract

[1] As of spring 2015, a new version of this database called GGOD has become available. For the purposes of this work, however, this is not relevant.

specific points-of-interest (POIs) such as educational or health facilities for each settlement.

Poland is divided into a hierarchy of administrative units. The top-level is formed by 16 administrative units called *voivodships*, with an average area of 20 000 km^2 (Central Statistical Office, 2015). The next lower level of the administrative hierarchy is the *district* level. There are 314 districts in Poland, with an average 20 districts per voivodship, and an average 994 km^2 per district.

For our experiments we selected 16 districts, representing approximately 5 % of the 314 Polish districts. In order to account for the variation in population density and settlement structure that can be found across the country, we selected districts from various parts of Poland and various voivodships that differ in terms of the structure of settlement size, settlement density, population density, the functional importance of settlements, and the topographic landscape. Figure 1 shows a map of population density in Poland with the selected districts highlighted.

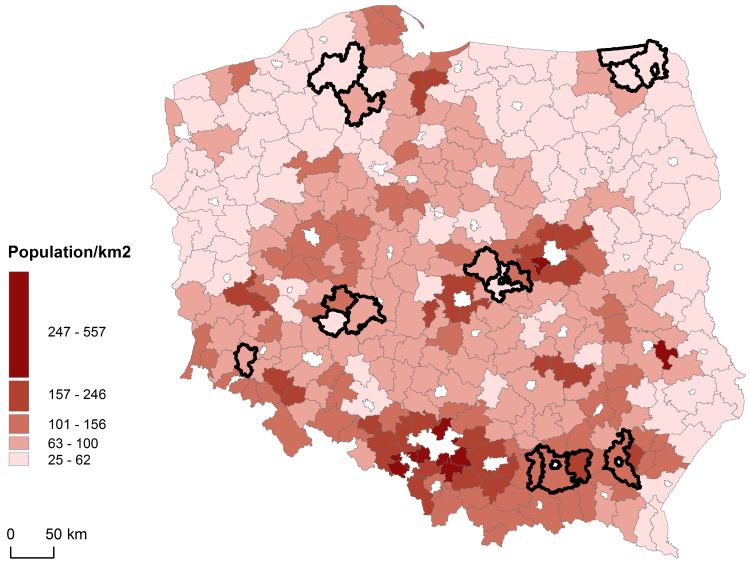

Figure 1. Choropleth map of population density per district in Poland, with the 16 selected districts highlighted (source: own elaboration, based on GGD data).

The 16 districts were grouped into 4 different types of settlement structure, as summarized in Table 1. Overall, 2 713 settlements were used in the experiments, out of 55 000 in all of Poland. This number is representative of the number of settlements in a typical Polish voivoidship (average per voivoidship is 3 400).

Table 1. Summary of districts used as study areas in the experiments.

Settlement structure	**Districts**	**Number of settlements**	**Characteristics**
HIGH_LARGE	Brzeski, Dębicki, Rzeszowski, Tarnowski (4)	720	high population density; settlements with large population
MED_MED	Krotoszyński, Ostrowski, Milicki, Złotoryjski (4)	473	medium to high population density; settlements with medium population
MED_SMALL	Łowicki, Skierniewicki, Żyrardowski (3)	538	medium population density; settlements with small population
LOW_SMALL	Bytowski, Chojnicki, Gołdapski, Olecki, Suwalski (5)	982	low population density; settlements with small population

Data enrichment

At the outset of this research the official set of rules published by GUGiK were analysed, governing the selection of settlements for the target scale of 1:500 000 (GUGiK, 2011). According to this rule set[2], settlements should be retained in the following cases:

[2] We use a slightly simplified version of the official rules, removing one special case that does not apply to the districts that form the basis of our experiments.

- settlements of type ‘city’
- settlements of type ‘village’ with municipal offices
- settlements of type ‘village’ with population > 100 habitants, within a district with population density < 50 inhabitants per km²
- settlements of type ‘village’ with population > 700 habitants, within a district with population density between 50 and 100 inhabitants per km²
- settlements of type ‘village’ with population > 1000 habitants, within a district with population density > 100 inhabitants per km²

From the above, it becomes obvious that only very basic properties of settlements are taken into account in the official rule set, which can be assigned to two groups of variables:

- thematic variables: population value, administrative status (seat of administrative office), settlement type (‘city’, ‘village’, ‘hamlet’ etc.)
- spatial variables: population density (per district)

However, as the analysis of published Polish map series shows, and as cartographic practice advises, other factors should also be taken into account, including the functional meaning of the settlements as well as the semantic and spatial relations among them. Thus, in order to reflect the settlements’ structure and their characteristics more thoroughly, additional variables should be exploited. However, these desired variables do not exist in the GGD database.

Owing to the semantic poverty of the GGD database the first stage of our methodology consisted in data enrichment, as a preprocessing step to the actual generalization process. Based on the study of manually generalized Polish paper maps as well as literature in urban geography (Carol, 1960; Smith, 1965; Batty, 2006) the

decision was made to use POI-related information, in particular buildings fulfilling specific functions from the BDOT10k database to enrich the GGD source database.

Using the basic ArcGIS and Python functionality the buildings representing relevant POI classes were selected from BDOT10k and spatially joined to the settlements in GGD. This allowed computing the number of POIs within the particular settlement belonging to one of 10 functional groups: health, communication and finance, accommodation, sacral, monumental, industrial, trading, educational, cultural, and other services (i.e. services not assigned to any other group, which might however be important for the selection process). This enrichment process made it possible to consider the importance of the settlements in GGD from the functional point of view, as trained cartographers would do in manual generalization.

The group of spatial variables was also supplemented, by calculating the value of the Voronoi area of each settlement as well as the distance to its nearest neighbor. In both cases, the centroid of the settlement polygon in GGD was used. Adding these two variables made it possible to better model the density of settlements and the distance relations among them. The enriched rule set was now able to make use of a total of 16 variables available for each settlement (new variables in italics):

- thematic variables: population, administrative status, settlement type, *functional significance (ten functional groups)*
- spatial variables: population density (per district), *Voronoi area, Nearest Neighbor distance*

In the experiments reported below, acronyms are used for these variables as shown in Table 2, which also lists the associated values.

Table 2. Variables with their acronyms and variable values used to characterize settlements.

Variable name	**Variable acronym**	**Variable values**
Population	POP	Last > 1500 Sixth 1001 - 1500 Fifth 501 - 1000 Fourth 201 - 500 Third 101 - 200 Second < 100
Administrative status	ADM	3 = district 4 = municipal -98 = none
Settlement type	TYP	96 = city 1 = village 2 = colony 3 = hamlet 4 = farmstead
Cultural function	Cult_f	0 - 4
Educational function	Edu_f	0 - 43
Trading function	Trade_f	0 - 697
Industrial function	Indust_f	0 - 526
Monumental function	Monum_f	0 - 3
Sacral function	Sacral_f	0 - 14
Accommodation	Accom_f	0 - 73
Communication and finance	Comm_f	0 - 31
Health function	Health_f	0 - 90
Other services	Other_f	0 - 35
Population density	POP_Dens	25.75 – 171.81
Voronoi area	V_Area	0.0000 - 38.3951
Nearest Neighbor distance	NEAR_DIST	0.0000 - 4698.7636

Machine Learning and Knowledge Extraction

In order to make use of the enriched GGD database, we formulated the settlement selection problem as a classification problem. Taking the enriched GGD database as a source, and taking the manually produced atlas map at the reduced scale of 1:500 000, (GGK, 1993-1997) as a target, we determined for all settlements whether they had been retained or removed at the target scale. We then added this status as a new variable to the attributes of the GGD attribute table. Thus, we had a classification problem with two labels (retained or removed), which could then be related to the thematic and spatial attributes of the enriched settlements layer of GGD. This allowed incorporating, in a reverse engineering approach, the cartographic knowledge that went into the manually generalized target map, stating for each settlement whether the trained cartographer who had made the map would select or omit it, depending on its functional properties.

We then developed two different settlement selection processes. The first one, termed the *basic approach* in the remainder of the paper, implements the rules of the official GUGiK documents (GUGiK, 2011), as specified in the preceding section, with the four variables population, population density, administrative status, and settlement type.

The second selection process, termed the *enhanced approach*, builds on a classification approach of the type described above. We used the set of 16 thematic and spatial variables described in the preceding section for that purpose. Since population density is computed at district level and thus is the same for all settlements of a particular district, we built a separate classification model for each of the four population density classes (HIGH_LARGE, MED_MED, MED_SMALL, LOW_SMALL, see Table 1), each time using the 16 settlement-specific variables as input features to the classification. The approach of separating into four different

population density classes follows the approach taken by GUGiK, who also distinguish between different population densities on a per-district basis. This solution is also justified by the results of earlier research on settlements selection problems (Karsznia 2013), which show that it is a reasonable approach to split the data and address generalization more locally than globally. In experiments not reported in this paper, we have also tried to build a single classification model based on the entire settlement dataset. However, the local solution with the split into population density classes resulted in richer and more meaningful DTs than the global approach. Thus, the density-classes approach was used in this paper. In order to build classification models, we used all settlements as input data and used 10-fold cross-validation to iteratively split the input data into a training and a testing subprocess (Geisser, 1993). The training subprocess is used for training a classification model, which is then applied in the testing subprocess. The model performance is evaluated during the testing phase. Within the cross-validation process the input data is randomly partitioned into subsets of equal size. From these subsets, a single subset is retained as the testing data set and constitutes an input for the testing subprocess, and the remaining subsets are used as training data set within a training subprocess. The cross-validation process is then repeated 10 times, with each of the 10 subsets used exactly once as the testing data set. The results from the iterations then can be averaged to produce a single estimate (Hasite et al., 2008) .

The classification models were implemented in RapidMiner 6, an open source machine learning and data mining software,[3] making use of three different ML algorithms (Rapid Miner User Manual , 2014; Rapid Miner Reference Manual, 2016):

[3] https://rapidminer.com/

- Decision trees (DT)
- Decision trees with optimized feature selection using a genetic algorithm (DT-GA)
- Support vector machines (SVM)

Machine Learning is a subfield of computer science and its objective is to pull the relevant information from the data and make it available to the user (Welling, 2010). There are two main approaches in machine learning. Supervised learning deals with predicting class labels from attributes using a labeled sample, unsupervised learning tries to discover interesting structure in the data. Here, we use the supervised learning approach. The aim of the supervised machine learning is to automatically extract rules from a set of given examples (called training set). The expert provides examples in the form of an object description using input attributes (called features), together with its corresponding classification (called labels). Machine-learning algorithms then automatically build rules from these. These rules can then be used to classify new examples provided to the system (Mustière, 2005). Among the classification methods available in the literature, decision trees are known not to deliver the best performance (Mitchell, 1997). However, they have two advantages that make them useful in the context of cartographic knowledge formalization. First, from the decision trees generated in the classification process, one can directly judge the importance of a particular input variable for the selection process. And second, DTs can be directly turned into human-readable rules. In the decision tree-based approach the goal is to create a classification model that predicts the value of a target attribute (the label) based on several input attributes (the features). Each interior node of a tree corresponds to one of the features. The number of edges of a nominal interior node is equal to the number of possible values of the corresponding feature. Outgoing edges of numerical attributes

are labeled with disjoint ranges. Each leaf node represents a value of the label attribute given the values of the input attributes represented by the path from the root to the leaf (Rapid Miner Reference Manual, 2016). The root of the tree shows the most important feature for the selection process (population in our case; which makes sense, as it is the most important variable in selecting settlements). The decision is made based on the terminal leaf and the path is directly represented on the tree.

The classification performance of DTs can be further enhanced if the selection of input variables (i.e. features) is optimized (Rapid Miner Reference Manual, 2016). For this purpose, we use a procedure implemented in RapidMiner based on a genetic algorithm (GA). Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. In the genetic algorithm used for feature selection 'mutation' means switching features on and off, while 'crossover' means interchanging the features used, thus leading to the selection of an optimal set of features (Rapid Miner Reference Manual, 2016).

As an alternative to DT and DT-GA, SVM was chosen as a state-of-the-art method that generally achieves strong classification performance (Welling, 2010). On the other hand, in contrast to DTs, the SVM method does not allow deriving the selection rules explicitly.

The evaluation of the results was carried out in different steps, and separately for each of the four types of settlement structures (HIGH_LARGE, MED_MED, MED_SMALL, LOW_SMALL) in the following way:

- Validation against the selection status acquired from the atlas map (taken as reference for evaluation).
- Comparison of performance statistics across the different approaches.

Results

Performance – basic approach

The performance achieved with the *basic approach* (using the GUGiK rules only) as compared to the manually generalized map can be found in Tables S1 to S4 (Supplementary Online Material**)**. The analyses were conducted within the previously defined district groups (cf. Table 1).

Performance – enhanced approach

The results of the performance of the *enhanced approach*, using the three different ML algorithms for the four district groups, are presented in Tables S5 to S8 (Supplementary Online Material**)**.

Performance – summary of results

As a summary of the tabular results, Table 3 shows the overall accuracy, as a global performance measure, for the basic approach and for the three methods of the enhanced approach, with best performance per district group highlighted.

Table 3. Overall performance for all methods. Bold typeface denotes best performance per district group. "Diff" denotes the difference between the best performing method and the basic approach.

District group	Overall accuracy in %				
	Basic approach	Enhanced approach			Diff
		DT	DT-GA	SVM	
HIGH_LARGE	83.38	85.14	**86.53**	86.11	3.15
MED_MED	77.59	80.75	**83.97**	80.96	6.38
MED_SMALL	81.78	82.53	85.15	**85.68**	4.90
LOW_SMALL	75.57	83.61	83.91	**87.38**	11.81

Decision trees

While the Tables S1 to S8 (Supplementary Online Material**)** and Table 3 allow to present the results in a compact but not spatially explicit way, the presentation of the results as maps and graphs takes up much more space, owing to the combinatorics of using 4 different district groups (each encompassing on average 4 districts) and 4 different methods (1 in the basic approach, 3 in the enhanced approach). We thus present only a fraction of the resulting maps and decision trees in the main body of this paper. For a more complete presentation of the results, we refer to the Supplementary Online Material ([insert URL to Figshare after review]). Figures 2 to 5 show examples of decision trees generated by the DT and the DT-GA method, respectively, for the district group with the smallest performance gain over the basic method (the group HIGH_LARGE) and for the district group with the largest performance gain (LOW_SMALL), according to Table 3. The acronyms as well as the variables' values shown on the decision trees correspond to the variables defined in Table 2. The labels on the arrows (i.e. on the links of the tree) correspond to the range of values defined or calculated for the particular variables. For instance, for the population (POP) variable six population classes had been defined, depending on the population count, labeled from 'second' (for the lowest population class) to 'last' (for the highest population). The definition of population classes was based on previous experience concerning the definition of class ranges for the purpose of small-scale maps (Karsznia 2013). Note, however, that the 'first' class was never used to form the decision trees shown in Figures 2 to 5. Similarly, the variable Cult_f (cultural function) in Figure 2 takes three discrete values (0, 1, 2), corresponding to the number of buildings fulfilling specific functions. Finally, for continuous variables such as NEAR_DIST or V_Area the link labels show the threshold which was calculated to further subdivide the tree. For instance, for the sixth POP class (Figure 2), on the second tree level a threshold value of NEAR_DIST of

2142.466 is used to split the remaining settlement objects into two groups. The boxes placed at the leaves of the DT show two aspects. First, the numbers shown indicate whether the corresponding tree path resulted in a selection [1.0] or omission [0.0] of the settlements concerned. Such a path can be short, as in the example of POP = fifth, which resulted in a selection; or such paths can extend over several levels, such as the path for POP = sixth AND NEAR_DIST ≤ 2142.466 AND Culf_f = 0.0 AND V_Area ≤ 2.131, which ultimately resulted in an omission (see Figure 2). The second aspect is shown as colored bars. It indicates the percentage of settlements meeting the particular condition, which were selected (blue) or omitted (red) in the training set. Hence, for POP = fifth about 60 % of settlements were selected, while about 40 % were omitted. Since the majority was selected, the numerical indicator turns to 1.

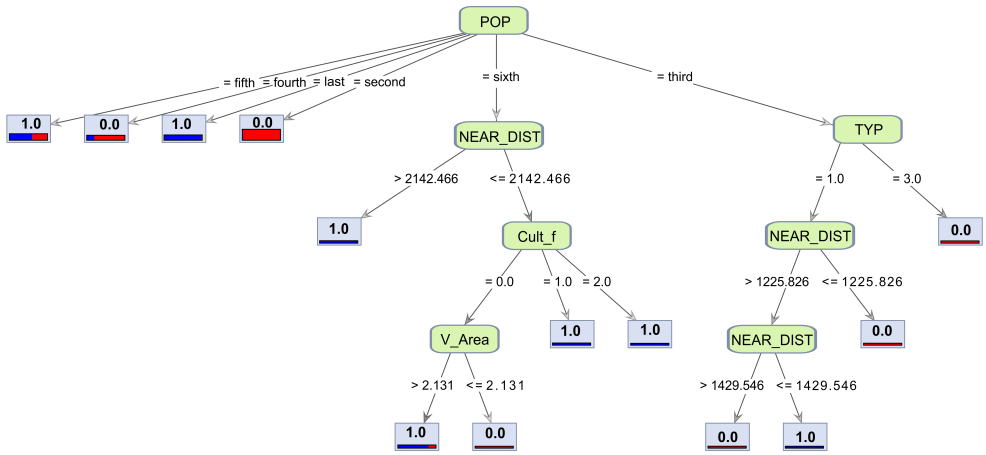

Figure 2. Decision tree generated by the DT algorithm for the HIGH_LARGE district group.

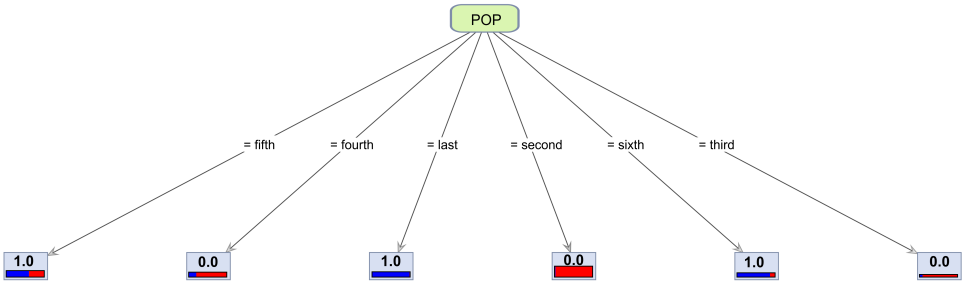

Figure 3. Decision tree generated by the DT-GA algorithm for the HIGH_LARGE district group.

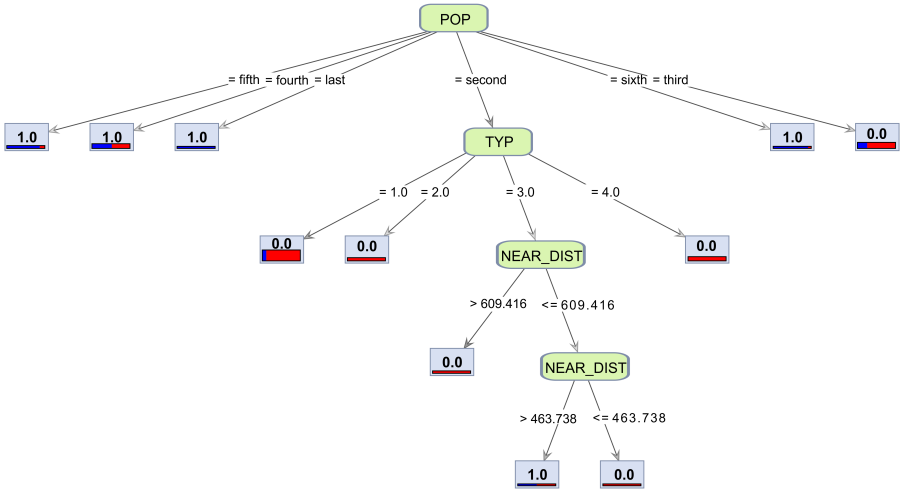

Figure 4. Decision tree generated by the DT algorithm for the LOW_SMALL district group.

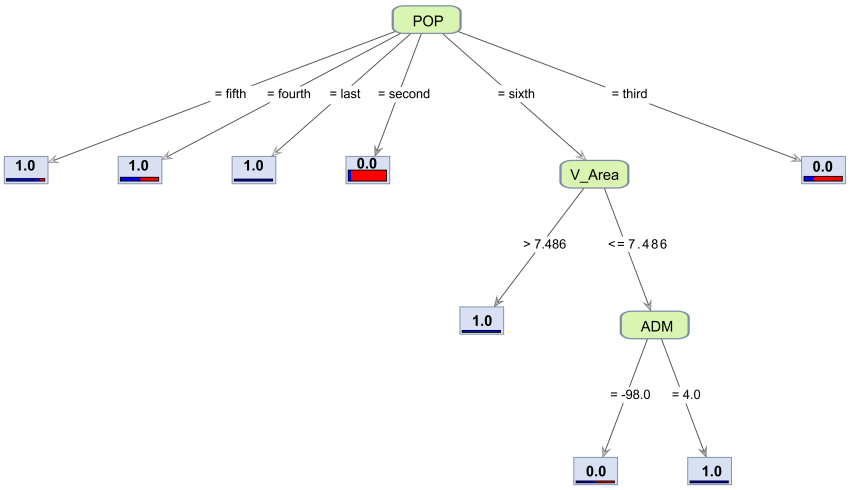

Figure 5. Decision tree generated by the DT-GA algorithm for the LOW_SMALL district group.

Map representation

As example maps, we have chosen three districts, shown in Figures 6 to 8. These examples have been chosen because they represent different district groups (MED_MED and LOW_SMALL) and because they show a range of different performances. Note that in all maps, the original road network from the GGD database is used, without generalization, owing to the fact that we focus exclusively on settlement selection in this paper.

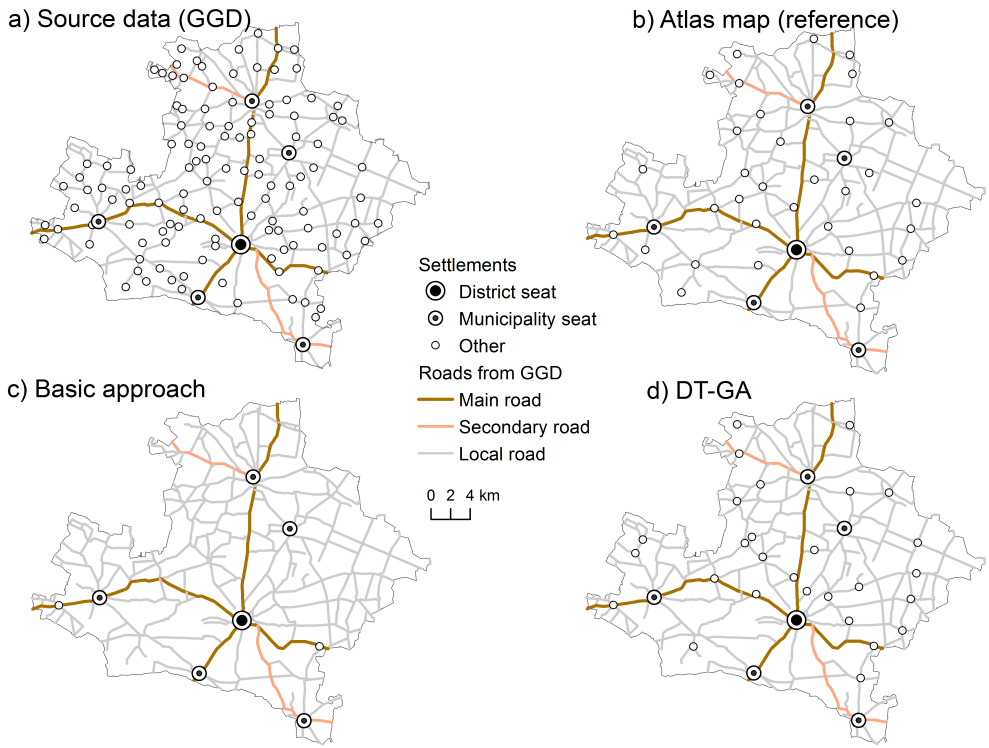

Figure 6. Maps of Krotoszyński district. (a) Source data in GGD. (b) Atlas map used as evaluation reference. (c) Basic approach. (d) DT-GA. Note that the original roads from GGD are used in all maps, i.e. they are not generalized.

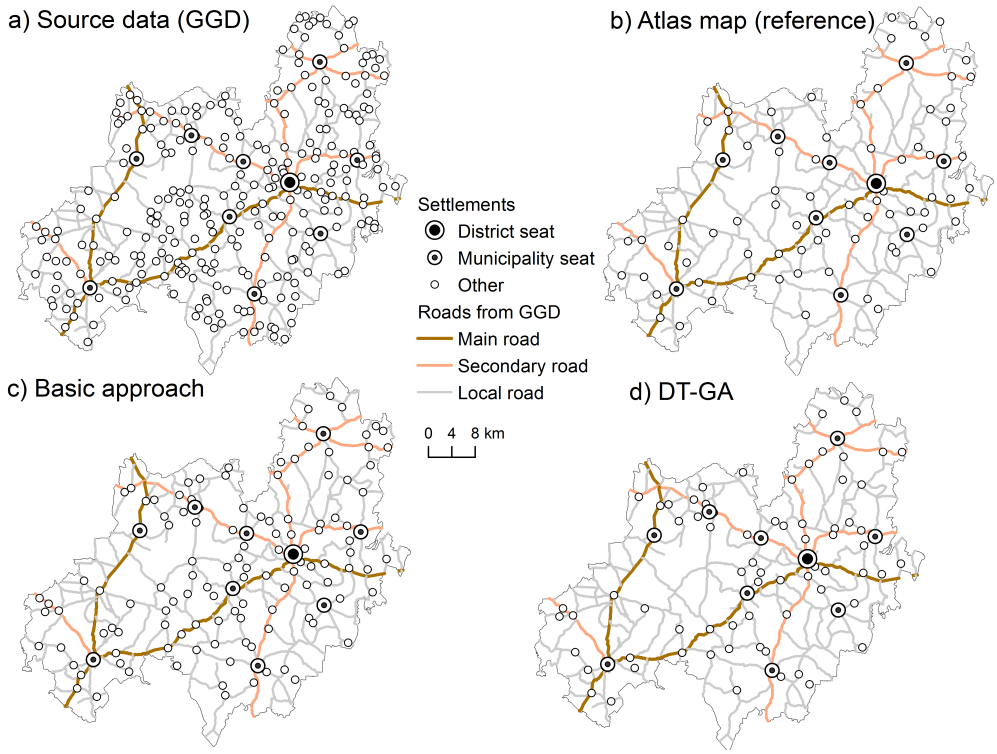

Figure 7. Maps of Bytowski district. (a) Source data in GGD. (b) Atlas map used as evaluation reference. (c) Basic approach. (d) DT-GA. Note that the original roads from GGD are used in all maps, i.e. they are not generalized.

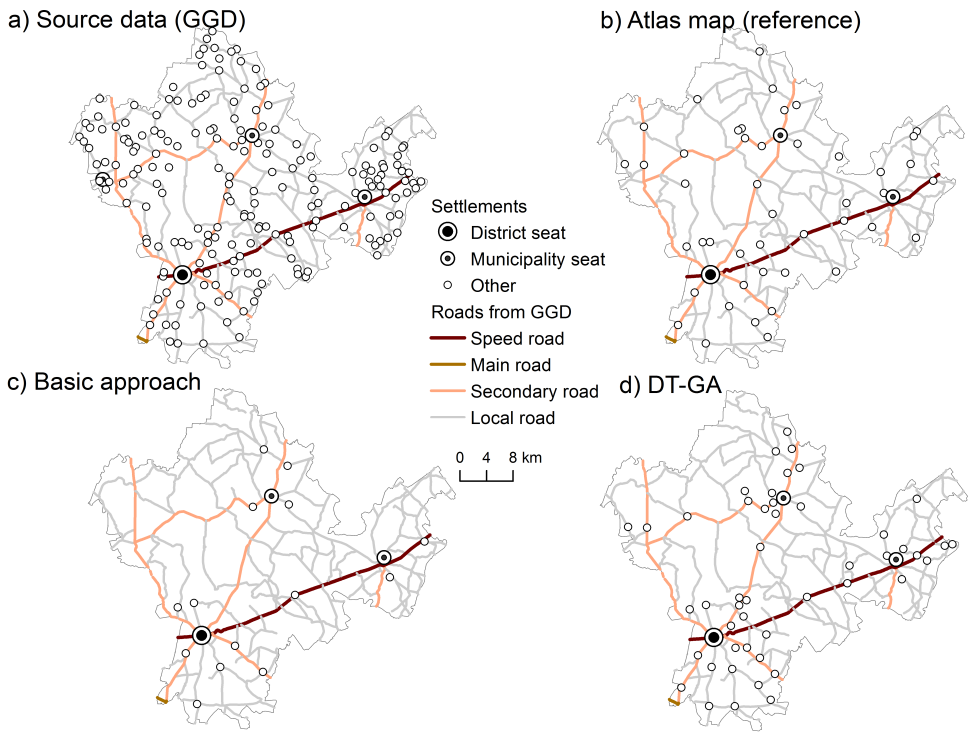

Figure 8. Maps of Chojnicki district. (a) Source data in GGD. (b) Atlas map used as evaluation reference. (c) Basic approach. (d) DT-GA. Note that the original roads from GGD are used in all maps, i.e. they are not generalized.

Discussion

With this work, we aimed to develop a methodology that combines data enrichment with machine learning to make settlement selection, as a subprocess of map generalization, more effective and more aware of semantic context. As stated in the Introduction, we aimed to make four contributions. The following discussion is therefore structured according to these claimed contributions.

Validation of written specifications and generalization rules

The validation of the official set of generalization rules against a manually produced map as well as the comparison between the *basic* and *enhanced approaches* suggests that the set of selection rules used by GUGiK does not cover the full range of settlement structures. In a manual production environment with trained cartographers, that is not a problem at all, since the cartographers will simply 'fill in' the gaps using their cartographic knowledge, based on their large experience of similar cases. However, in an automated system, the rule set needs to be as complete as possible. The proposed methods of this paper help developing such more holistic rule sets, and they are sufficiently generic such that they can be transposed to the context of other NMAs. This particular step may not be a very spectacular contribution. However, we believe it is nevertheless useful, as we have been able to pinpoint shortcomings in the currently used rule set for settlement selection, and we have demonstrated that it can actually be improved and extended.

Data enrichment

The data enrichment step has benefits in three respects. First, adding more semantics to the data makes the description of settlements more complete. Second, the functional importance of the settlements makes the generalization process more context-dependent and more similar to the manual process carried out by trained cartographers. Third, it can help discovering characteristic functional patterns in the data using machine learning.

In order to take into account, the variation in the settlements' spatial structure, and thus introduce context-dependence, we have introduced two new variables to the basic GUGiK variables: Voronoi area (V_Area) and nearest neighbour distance (NEAR_DIST). In the *enhanced approach* both variables proved to be decisive in the

settlement selection process, as they both appeared in the decision trees obtained within the classification process for the HIGH_LARGE and LOW_SMALL districts groups.

Machine learning

The idea behind using machine learning algorithms was to reveal potential additional selection rules, thus making explicit the experience of trained cartographers implicitly contained in published atlas maps. Hence, the *basic approach*, implementing the written rules of GUGiK, was compared to three different ML algorithms of the *enhanced approach* (DT, DT-GA, SVM). As a reference data set for the evaluation of the results generated by both the basic and the enhanced approach, as well as for training the ML algorithms, we used a manually produced atlas map at the target scale of 1:500 000 (GGK 1993-1997). That is, we assume the selection of settlements in the atlas map to be perfect. Naturally, that assumption is debatable, as it is known that even the generalization results of experienced cartographers may differ (Duchêne et al. 2014), and hence there is probably no single 'correct' solution. This points to a common problem in the evaluation of generalization results: What is the reference to compare to? And it points to a key limitation of supervised ML approaches: The results are constrained by the quality of the training data set used (apart from a number of other factors). The problem could to some extent be reduced by using several reference sources and integrating these to an 'average' reference. However, since only a single manually produced map was available, we will use that single source (GGK 1993-1997). Also, from a methodological perspective, the fairness of the evaluation procedure is guaranteed, as we use the same reference for both the basic and the enhanced approach.

While the detailed evaluation results for the basic approach are provided in Tables S1 to S4, and the results of the evaluation of the enhanced approach are shown in

Tables S5 to S8 (Supplementary Online Material**)**, we focus our discussion on Table 3, which summarizes and compares the accuracy of settlement selection across all district groups and selection approaches used. From this table, we gain several insights. First, the enhanced approach always performs better than the basic approach, ranging from a mere 3.15 % improved selection accuracy in the HIGH_LARGE districts to almost 12 % in the LOW_SMALL districts (shown in the Diff column), suggesting that the basic set of selection rules may have particular deficiencies in rural areas. Second, even the simplest ML classifier used, decision trees (DT), consistently outperforms the basic approach. Third, the DT-GA and the SVM method, respectively, perform even better. For each of the four district groups, the accuracy of the simple DT is about halfway between the accuracy of the basic approach and the best results obtained. Fourth, except for the LOW_SMALL districts, the performance of the DT-GA is equal or better than the accuracy of the SVM classifier, suggesting that the GA-based optimization of classification feature selection brought the results to the same level as the more sophisticated SVM method. This is important for the translation of our ML results into selection rules, as DTs are much more easily translated into rules than SVM results (cf. the discussion in the following subsection). Fifth, and finally, with best values for selection accuracy ranging between 83.97 % and 87.38 % (the bold numbers in Table 3), we are probably starting to approach the quality level of different manually produced maps, originating from different authors. As mentioned above, different authors will probably produce different solutions, and the variation of these differences may easily amount to a few percent. Hence, we probably do not need to reach 100 % selection accuracy to be 'good enough'.

What exactly led to the improvement that we have seen for the *enhanced approach*? Figures 2 to 5, showing four examples of decision trees generated, give some

insight on this. More decision trees are included in the Supplementary Online Material (Figures S1 to S4). Basically, these trees reveal two effects. First, we see that the POP variable, which had already been included in the written rules of the basic approach, remains most decisive: POP appears close to the root of the tree, and it dominates the trees. Nevertheless, we also see that the new variables introduced in the data enrichment step, such as NEAR_DIST, Cult_f, and V_Area are now integrated into the decision trees (Fig. 2, Fig. 4, Fig. 5 ; Fig. S1 and S2). Since these variables did not exist in the basic approach and since the enhanced approach consistently outperformed the basic approach, we conclude that these enriched semantics — in combination with a more detailed and differentiated setting of class boundaries for the POP variable — made the difference in selection performance. The second effect that becomes noticeable is the impact of the optimized set of classification features used in DT-GA, as opposed to simple DT. This effect can be seen when comparing Figure 2 with Figure 3, and Figure 4 with Figure 5, respectively. Each of these two pairs of figures shows the decision tree for simple DT vs. DT-GA for one particular group of districts. In both cases (Fig. 2 vs. Fig. 3 and Fig. 4 vs. Fig. 5), we see that GA-based optimization has led to simpler decision trees. In the first pair (Fig. 2 vs. Fig 3), the decision tree is reduced from four levels to only one level. In the second pair, the decision tree was reduced from four levels (Fig. 4) to three levels (Fig. 5), again introducing variables from the enriched variable set (V_Area). Even if multiple variables are introduced to the decision tree, the GA optimization leads to simplified trees. We have observed this complexity reduction in most cases, but there are also counter-examples, as shown in Figures S1-S4.

Finally, we take a look at the cartographic visualization of the settlement selection process. The maps shown in Figures 6 to 8 represent a small sample of

districts; more maps can be found in the Supplementary Online Material (Figures S5 and S6). Figure 6 displays the resulting map for the Krotoszyński district, which is a representative of the MED_MED districts, the group that showed the second-highest improvement over the basic approach (6.38 %). We can see that the map generated by the DT-GA method clearly matches the atlas map better than the map resulting from the basic approach. This is confirmed by the numerical performance measures (82.5 % overall accuracy for DT-GA, as opposed to 75.2 % for the basic approach). The Figures 7 and 8, respectively, show two representatives of the LOW_SMALL districts, the group that scored the highest improvement in the enhanced approach (almost 12 %). In Figure 7, for the Bytowski district, we see a generally good agreement of the DT-GA method with the atlas map, and a slightly better performance than for the basic approach (85.07 % overall accuracy for DT-GA vs. 77.9 % for the basic approach). Similarly, in the map of the Chojnicki district (Fig. 8), we can see that the DT-GA performs better than the basic approach (DT-GA: 90.86 % overall accuracy; basic approach: 83.25 %). The LOW_SMALL district group is characterized by many small settlements that all have a similar population and, being small villages or hamlets, have no particular distinctive functions. Thus, the spatial variables obtain a key role, since the spatial density and arrangement of settlements becomes the main driver for settlement selection. Both newly introduced spatial variables NEAR_DIST and V_Area, which express the spatial relations required in the selection process appear on the decision tree created for the Chojnicki district (see Figure S4), however, they are not used in the decision tree for the whole LOW_SMALL group (Fig. 5).

Translating ML results to selection rules

Usually the basic generalization rules are given in the form of written specifications and official documents, such as GUGiK (2011). However, they are seldom very detailed and

formalized (Müller and Mouwes, 1990). In order to build the detailed cartographic knowledge base the procedural knowledge hidden in well-design maps has to be explored and made explicit. In this paper we demonstrate that appropriate ML models can offer an excellent tool to discover generalization rules. Thanks to classification models using decision trees we have identified new variables that are decisive in the settlement selection process. Moreover, from the generated decision trees, we can directly infer generalization rules in the form of if-then rules. For instance, based on the DT- GA tree obtained for the LOW_SMALL district group we can formulate the following settlement selection rule (see Figure 9): IF POP in range = 'sixth' AND V_Area < = 7.486 AND ADM = 4.0 THEN the settlement should be selected (because the settlement status equals 1.0, that is, 'selected').

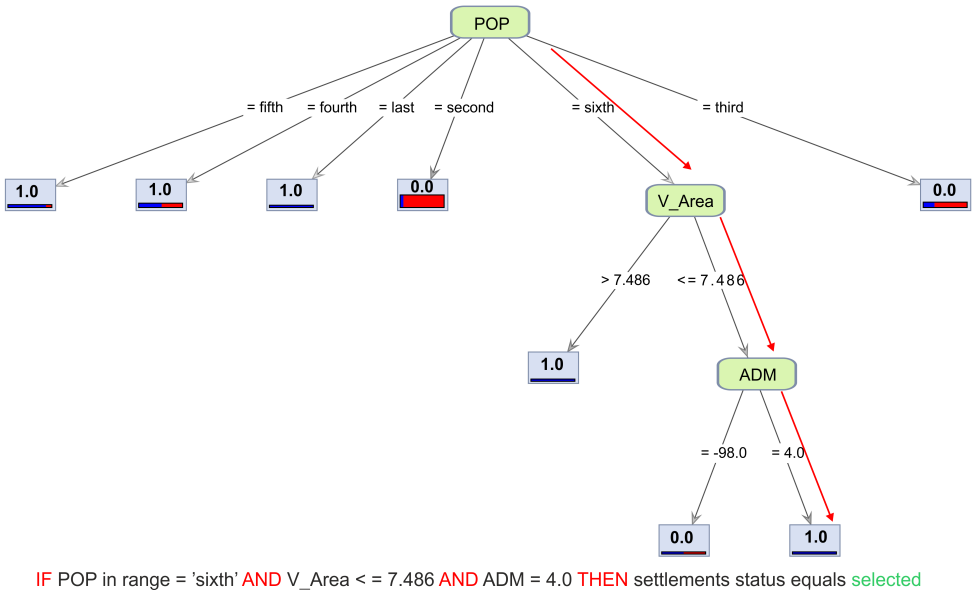

Figure 9. Example of a settlement selection rule read directly from a decision tree.

Other classification algorithms, such as SVM, random forests, or neural networks, may generate better classification results. However, none of these algorithms delivers results that can so easily be interpreted and transformed into human-readable results — which is of key importance when the ultimate aim is to support rule formation

for cartographic practice. Additionally, in our experiments DT in combination with optimized feature selection was always close to SVM in performance, or even outperformed SVM.

Conclusions

In this paper, we have used the example of settlement selection for small-scale maps to show how written map specification and generalization rules in use at NMAs may be incomplete, how these knowledge gaps may be discovered, and how more complete and context-dependent rules may be obtained by first enriching an existing topographic database with additional semantics and then using machine learning tools to extract new rules automatically from the data. As discussed above in detail, this work has made methodological contributions in four respects: regarding the validation of written specifications and generalization rules; a possible approach for data enrichment; a methodology for using machine learning tools in knowledge formalization; and regarding the transformation of machine learning results into human-readable cartographic rules. The fact that we do not reach 100% classification accuracy means two things: First, there is probably still more 'deep knowledge' knowledge to be discovered, possibly linked to further variables that we did not include in the data enrichment, such as the connectivity of the road network: small settlements are for instance often chosen when they are located at nodes of degree ≥ 3. Second, the solutions of manual cartography, which were taken as 'ground truth' are known to vary, and hence a 'perfect' result cannot be obtained.

In future work it would be interesting to use the proposed ML methodology to compare the generalization rules that exist at different NMAs in order to attempt extracting general, common rules for small-scale map generalization. This approach might also lead to discovering potential cultural differences between considered NMAs,

countries, or even ‘cartographic schools’. This could also complement the research conducted for the large-scale maps by Stoter et al. (2009).

Since high-resolution data with detailed attribute information is not available from official sources for all parts of the world, exploiting user-generated data such as OpenStreetMap for the data enrichment process could also be explored. User-generated data is known to commonly contain more semantic attributes, and hence might bear even better potential for semantic enrichment than topographic data from official sources. Also, user-generated data represent the *vernacular*, rather than the official, view of geography; this could be of interest when vernacular maps should be produced (Aliakbarian and Weibel 2016). On the other hand, such data sources, though available for the entire world, may introduce new challenges. Research has, for instance, highlighted the varying degree of detail with which OpenStreetMap data has been captured (Girres and Touya 2010; Touya and Brando 2013). The use of data sources such as OpenStreetMap in cartography will therefore provide an interesting field of future research.

Finally, the conducted research would also require to be extended to integrate other thematic layers in the generalization process such as the road network or the river network in order to achieve more holistic solutions. The proposed approach is however a first step towards a comprehensive methodology and toolbox for small-scale map generalization.

Acknowledgements

The Authors would like to express their gratitude to SCIEX SCIentific EXchange program, which by supporting the effective scientist’s mobility, made possible to conduct this research.

The Authors also gratefully acknowledge Professor Wieslaw Ostrowski as well as M.Sc Ali Soleymani for their contributions to the discussions of this project.

References

Ai T., and Liu Y. 2004. Analysis and simplification of point cluster based on Delaunay triangulation model. In Z. Li, Q. Zhou and W. Kainz (Eds.), *Advances in spatial analysis and decision making* (pp. 9-19), London: Taylor & Francis.

Aliakbarian, M. and Weibel, R. 2016. Integration of folksonomies into the process of map generalization, *ICA Workshop on Generalization and Multiple Representation*, Helsinki, Finland. http://tinyurl.com/hyttlzk

Armstrong M. P., 1991. Knowledge classification and organization. In B. Buttenfield and R. McMaster (Eds.), *Map generalization: making rules for knowledge representation* (pp. 86-102), London: Longman.

Balley S., Baella B., Christophe S., Pla M., Regnauld N. and Stoter J. 2014. Map Specifications and User Requirements. In D. Burghardt, C. Duchêne and W. Mackaness (Eds.) *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation* (pp. 17-52), Cham, Switzerland: Springer.

Batty M. 2006. Hierarchy in cities and city systems. In D. Pumain (Ed.) *Hierarchy in Natural and Social Sciences* (pp. 143-168), Netherlands: Springer.

Bobzien, M., Burghardt, D., Neun, M. and Weibel, R. 2008. Multi-Representation Databases with Explicitly Modeled Horizontal, Vertical and Update Relations. *Cartography and Geographic Information Science*, 35(1): 3-16. DOI: 10.1559/152304008783475698

Brassel, K. E., and Weibel, R. 1988. A Review and Conceptual Framework of Automated Map Generalization, *International Journal of Geographical Information Systems*, 2(3): 229–244. DOI: 10.1080/02693798808927898

Burghardt, D., Duchêne, C. and Mackaness, W.A. 2014. *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation*. Cham, Switzerland: Springer.

Buttenfield B. P., McMaster R. B. 1991. *Map generalization: Making rules for knowledge representation.* Harlow: Longman.

Carol H. 1960. The hierarchy of central functions within the city. *Annals of the Association of American Geographers*, 50(4): 419-438. DOI: 10.1111/j.1467-8306.1960.tb00359

Central Statistical Office. 2015. *Area and population in the territorial profile in 2015: Statistical information and elaborations*, Warsaw. http://tinyurl.com/ntcqssv

Chang, H. and McMaster, R. B. 1993. Interface design and knowledge acquisition for cartographic generalization. In *Proceedings AutoCarto 9*. Bethesda, ACSM/ASPRS: 187–96. http://tinyurl.com/pc6skh5

Dixon O. M. 1967. The Selection of Towns and Other Features on Atlas Maps of Nigeria. *The Cartographic Journal*, 4(1): 16-23. DOI: 10.1179/caj.1967.4.1.16

Duchêne C., Dadou M., Ruas A. 2005. Helping the capture of expert knowledge to support generalization. *ICA Workshop on Generalization and Multiple Representation*, A Coruña, Spain. http://tinyurl.com/plsum7c

Duchêne C, Baella B., Brewer C. A., Burghardt D., Buttenfield B. P., Gaffuri J., Käuferle D.,Lecordix F., Maugeais E., Nijhuis R., Pla M., Post M., Regnauld N., Stanislawski L. V., Stoter J., Tóth K., Urbanke S., van Altena V. and Wiedemann A. 2014. Generalisation in Practice Within National Mapping Agencies. In D. Burghardt, C. Duchêne and W. Mackaness (Eds.) *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation* (pp. 329-391), Cham, Switzerland: Springer.

Dutton, G. and Edwardes, A.J. 2006. Ontological Modeling of Geographical Relationships for Map Generalization. *ICA Workshop on Generalization and Multiple Representation*, Portland, OR. http://tinyurl.com/njtlkzf

Flewelling D. M., Egenhofer M. J. 1993. Formalizing importance: parameters for settlement selection from a geographic database, *Proceedings of Auto-Carto XI*, Minneapolis. http://tinyurl.com/jabyyuv

Geisser, S. 1993. *Predictive Inference*. New York: Chapman and Hall.

GGK. 1993–1997. Atlas Rzeczypospolitej Polskiej. Mapa 1:500 000, Główny Geodeta Kraju, Warszawa,. [Atlas of the Republic Poland, 1: 500 000 map, The Surveyor General of Poland, Warsaw, 1993-1997])

Girres, J.-F. and Touya G. 2010. Quality Assessment of the French OpenStreetMap Dataset, *Transactions in GIS*, 14(4): 435-460. DOI: 10.1111/j.1467-9671.2010.01203

GUGiK. 2011, Regulation of the Ministry of the Interior from 17 of November 2011 on the topographic objects database and general objects database as well as standard cartographic products, *Head Office of Geodesy and Cartography guidelines*. http://tinyurl.com/p5r3b46

Harrie L., Weibel R. 2007. Modelling the Overall Process of Generalisation. In W.A. Mackaness, A. Ruas and L.T. Sarjakoski (Eds.) *Generalisation of Geographic Information: Cartographic Modelling and Applications* (pp. 67-87), Oxford: Elsevier.

Hastie T., Tibshirani R., Friedman J. 2008. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Second Edition, Springer.

Kadmon N. 1972. Automated Selection of Settlements in Map Generalisation. *The Cartographic Journal,* 9(2): 93–98.

Karsznia I. 2013. Selected aspects of settlement generalization in the General Geographic Database in Poland. *Geodesy and Cartography*, 39(3): 129-137.

Kilpeläinen T. 2000. Knowledge Acquisition for Generalization Rules, *Cartography and Geographic Information Science*, 27(1): 41-50. DOI: 10.1559/152304000783547993

Kulik, L., Duckham, M. and Egenhofer, M. 2005. Ontology-driven map generalization. *Journal of Visual Languages & Computing*. 16(3): 245-267, DOI: 10.1016/j.jvlc.2005.02.001

Langran C. N, Poiker T. K. 1986. Integration of name selection and name placement. *Proceedings of 2nd International Symposium on Spatial Data Handling*: 50-64.

Leitner, M. and Buttenfield, B.P. 1995. Acquisition of procedural cartographic knowledge by reverse engineering. *Cartography and Geographic Information Systems*, 22(3): 232–241.

Li, Z. 2007. *Algorithmic Foundation of Multi-Scale Spatial Representation*. CRC Press.

Mackaness, W.A., Ruas, A. and Sarjakoski, L.T. 2007. *Generalisation of Geographic Information: Cartographic Modelling and Applications,* Oxford: Elsevier.

Mackaness W. A., Gould N. M. 2014a. The role of Geography in Automated Generalization. *ICA Workshop on Generalization and Multiple Representation*, Vienna, Austria. http://tinyurl.com/hozn5tj

Mackaness, W.A., Burghardt, D. and Duchêne, C. 2014b. Map Generalisation: Fundamental to the Modelling and Understanding of Geographic Space. In D. Burghardt, C. Duchêne and W. Mackaness (Eds.) *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation* (pp. 1-15), Cham, Switzerland: Springer.

Mitchell T. M. 1997. *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Müller, J.C. and Mouwes, PJ. 1990. Knowledge Acquisition and Representation for Rule Based Map Generalization: An Example from the Netherlands, *Proceedings of GIS/LIS 90*, Anaheim, CA: 58-67.

Müller J. C., Lagrange J. P., and Weibel R. 1995a. *GIS and Generalization. Methodology and Practice*, Taylor & Francis.

Müller J. C., Weibel R., Lagrange J. P., Salgé F. 1995b. Generalization: state of the art and issues. In J. C. Müller, J. P. Lagrange and R. Weibel (Eds.) *GIS and generalization. Methodology and practice* (pp. 3-17), Taylor & Francis.

Mustière, S., Zucker, J.-D. and Saitta, L. 2000. An abstraction-based machine learning approach to cartographic generalization. In *Proceedings of 9th international symposium on spatial data handling*. Beijing, sec. 1a, 50–63.

Mustière, S. 2005. Cartographic generalization of roads in a local and adaptive approach: A knowledge acquisition problem. *Int. Journal of Geographical Information Science*, 19(8-9): 937-955. DOI: 10.1080/13658810509161245

Neun M. 2007. *Data enrichment for adaptive map generalization using web services*. Doctoral thesis, Department of Geography, University of Zurich. http://tinyurl.com/ohhblh6

Neun, M., Burghardt, D. and Weibel, R. 2008. Web Service Approaches for Providing Structural Cartographic Knowledge to Generalisation Operators. *Int. Journal of Geographical Information Science*, 22(2): 133-165. DOI : 10.1080/13658810701348997

Plazanet C., Bigolin N. M., Ruas A. 1998. Experiments with Learning Techniques for Spatial Model Enrichment and Line Generalization . *GeoInformatica International Journal Issue* 2(4): 315-333, The Netherlands, Kluwer Academic Publishers. DOI: 10.1023/A:1009753320636.

Rapid Miner Reference Manual, 2016. http://docs.rapidminer.com/studio/operators/rapidminer-studio-operator-reference.pdf (access: 03.08.2016)

Rapid Miner User Manual. 2014. http://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf (access: 03.08.2016)

Ratajski L. 1973. Considerations in cartographic generalization – in Polish (Rozważania generalizacji kartograficznej). *Polish Cartographical Review*, 5(2): 49–54 (Part I) ; 5(3): 103–110 (Part II).

Richardson D. E., Müller J-C. 1991. Rule selection for small-scale map generalization. In B. Buttenfield, R. McMaster (Eds.) *Map generalization: making rules for knowledge representation* (pp. 136–149), London: Longman.

Ruas, A., Dyèvre A., Duchêne C and Taillandier P. 2006. Methods for improving and updating the knowledge of a generalization system, *16th AutoCarto Research Symposium*.

Samsonov T. E, Krivosheina A. M. 2012. Joint generalization of city points and road network for small-scale mapping*, GIScience conference proceedings*, Columbus, Ohio. http://tinyurl.com/p8t7fk9

Sheeren D., Mustière S. and Zucker, J.-D. 2009. A data-mining approach for assessing consistency between multiple representations in spatial databases. *International Journal of Geographical Information Science*, 23(8): 961 – 992. DOI:10.1080/13658810701791949

Smith R. H. T. 1965. Method and Purpose in Functional Town Classification. *Annals of the Association of American Geographers*, 55(3): 539-548.

Stanislawski L. V., Buttenfield B. P., Bereuter P., Savino S., Brewer C. A. 2014. Generalization Operators. In D. Burghardt, C. Duchêne and W. Mackaness (Eds.) *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation* (pp. 157-196), Cham, Switzerland: Springer.

Steiniger S., Weibel R. 2007. Relations Among Map Objects in Cartographic Generalization. *Cartography and Geographic Information Science*, 34(3): 175-197.

Steiniger S., Taillandier P. and Weibel R. 2010. Utilising urban context recognition and machine learning to improve the generalisation of buildings, *International Journal of Geographical Information Science,* 24(2): 253-282. DOI: 10.1080/13658810902798099

Stoter J, van Smaalen J., Bakker N., Hardy P. 2009. Specifying Map Requirements for Automated Generalization of Topographic Data. *The Cartographic Journal* 46(3): 214–227. DOI: 10.1179/174327709X446637

Stoter J, Baella B, Blok C, Burghardt D, Duchêne C, Pla M, Regnauld N, Touya G. 2010. State of the art of automated generalisation in commercial software. EuroSDR official publication no. 58. http://tinyurl.com/nkeutaj

Stoter J., Post M., van Altena V., Nijhuis R. & Bruns B. 2014. Fully automated

generalization of a 1:50k map from 1:10k data, *Cartography and Geographic Information Science*, 41(1): 1-13. DOI: 10.1080/15230406.2013.824637

Touya, G. and Brando C. 2013. Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets, *Cartographica*, 48(2): 134-143. DOI: 10.3138/carto.48.2.1836

Töpfer F., Pillewizer W. (1966). The Principles of Selection. *The Cartographic Journal*, 3(1): 10-16. DOI: 10.1179/caj.1966.3.1.10

Weibel R. 1995a. Three essential building blocks for automated generalization. In J. C. Müller, J. P. Lagrange and R. Weibel (Eds.) *GIS and Generalization. Methodology and Practice*, (pp. 3-17), Taylor & Francis.

Weibel R., Keller S. et Reichenbacher T. 1995b. Overcoming the Knowledge Acquisition Bottleneck in Map Generalization: the Role of Interactive Systems and Computational Intelligence. *Lecture Notes in Computer Science*. Proceedings COSIT '95, (Vol. 988): 139-156, Berlin: Springer-Verlag.

Welling M., 2010. A First Encounter with Machine Learning. Donald Bren School of Information and Computer Science. University of California Irvine.

West-Nielsen P., Meyer M. 2007. Automated Generalisation in a Map Production Environment – the KMS Experience. In W.A. Mackaness, A. Ruas and L.T. Sarjakoski (Eds.) *Generalisation of Geographic Information: Cartographic Modelling and Applications*, (pp. 301-313), Elsevier.

Van Kreveld M., Van Oostrum R., Snoeyink J. 1997. Efficient settlement selection for interactive display. Proceedings Auto-Carto XIII, Seattle, USA.

Yan H., Weibel R. 2008. An algorithm for point cluster generalization based on the Voronoi diagram. *Computers and Geosciences*, 34(8): 939-954. DOI: 10.1016/j.cageo.2007.07.008